University of Antwerp
Operations Research Group
ANT/OR

Dance Hit Prediction

Dorien Herremans, David Martens, Kenneth Sörensen

Workshop on Music and Machine Learning
Prague, 23.09.2013

Overview

- Problem description
- Database
- Preprocessing
- Models
- Results
- Conclusion

Top 10 dance hit?

"Pachet, François, and Pierre Roy. "Hit song science is not yet a science." Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008). 2008."

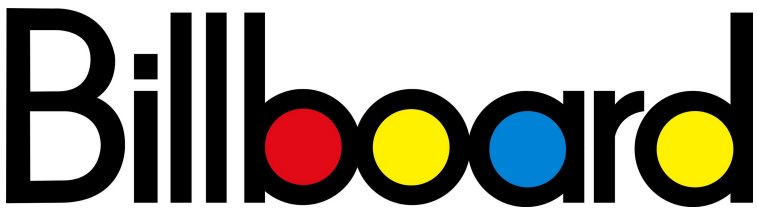
- Record companies
- Classification problem

Data Sources

- Official Charts company (OCC) - officialcharts.com
  - Weekly top 40 dance hits
  - 10/2009 until 3/2013
  - 7159 hits (759 unique songs)
  - 7090 with data (697 unique songs)

Data Sources

- Billboard (BB) - billboard.com
  - Weekly top 10 dance hits
  - 1/1985 until 3/2013
  - 14533 hits (3361 unique songs)
  - 12711 with data (2755 unique songs)



Extracting Hit Data

- Step 1: Download HTML pages
  - HTML document with all download links . . .
  - Firefox: DownThemAll

```
<html><a href=" http://www.billboard.com/charts/ 2013-03-02 /dance-club-play-songs ">1</a>
<a href=" http://www.billboard.com/charts/ 2013-02-23 /dance-club-play-songs ">1</a>
<a href=" http://www.billboard.com/charts/ 2013-02-16 /dance-club-play-songs ">1</a>
...
```

Extracting Hit Data

- Step 2: Parse information
  - JSoup parser: Open source Java HTML parser
  - Extract: song title, artist, position, date

```
Elements parents = doc.select("article.song_review");
for (Element parent : parents)
{
    Elements position = parent.select("span.chart_position");
    Elements titles = parent.select("h1");
    Elements categories = parent.select("p.chart_info a");
    Elements catnolink = parent.select("p.chart_info"); //to correct if there is no link tag around
        out.write("date - " + date.text() + " - position - " + position.text() + " - song - " + titles.text() + " - ");
        if (categories.text().length()<1){
                out.write("artist - " + catnolink.text() + "\n");
        }else {
                out.write("artist - " + categories.text() + "\n");
        }
}
```

Getting Music Information

- The Echo Nest:
  - World's leading music intelligence company
  - Largest repository of dynamic music data in the world
  - Over a trillion data points on over 30 million songs
  - Clear Channel, EMI, eMusic, MOG, MTV, Nokia, Rdio, Spotify, Twitter, ...

Getting Music Information

- Echo Nest Java API (jEN): open source Java client library for the Echo Nest developer API.
- Extract and calculate musical features for hits
- Challenges: Featuring, Ft, Feat, (), /, Remix, With...

```
if (result.get(0)=="not found"){
    arr = hits.get(k).get(5).split(" Ft");
    result = sse.findSongByTitle(hits.get(k).get(4).replaceAll("\\(.*\\)", ""), arr[0], 1);
```

Extracted Features

- 139 usable attributes
- Meta data (discarded)
  - Artist
  - Song
  - Artist Location
  - Artist Familiarity
  - Artist Hottness
  - Song Hottness

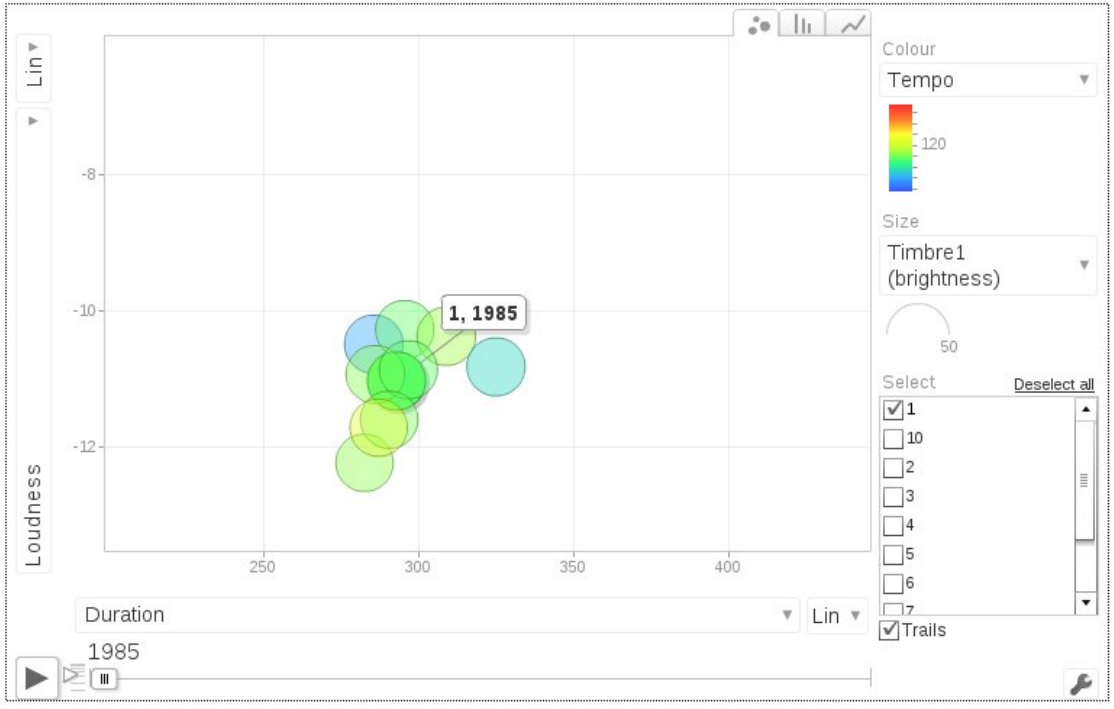
Extracted Features

- Analyser data
  - Duration (seconds)
  - Tempo (beats per minute)
  - Time signature: how many beats in one bar
  - Mode (minor - major)
  - Key: estimated overall key of a track
  - Loudness: average (dB)
  - Energy: mix of loudness and segment durations
  - Danceability: mix of beat strength, tempo stability, overall tempo,. . .

Extracted Features

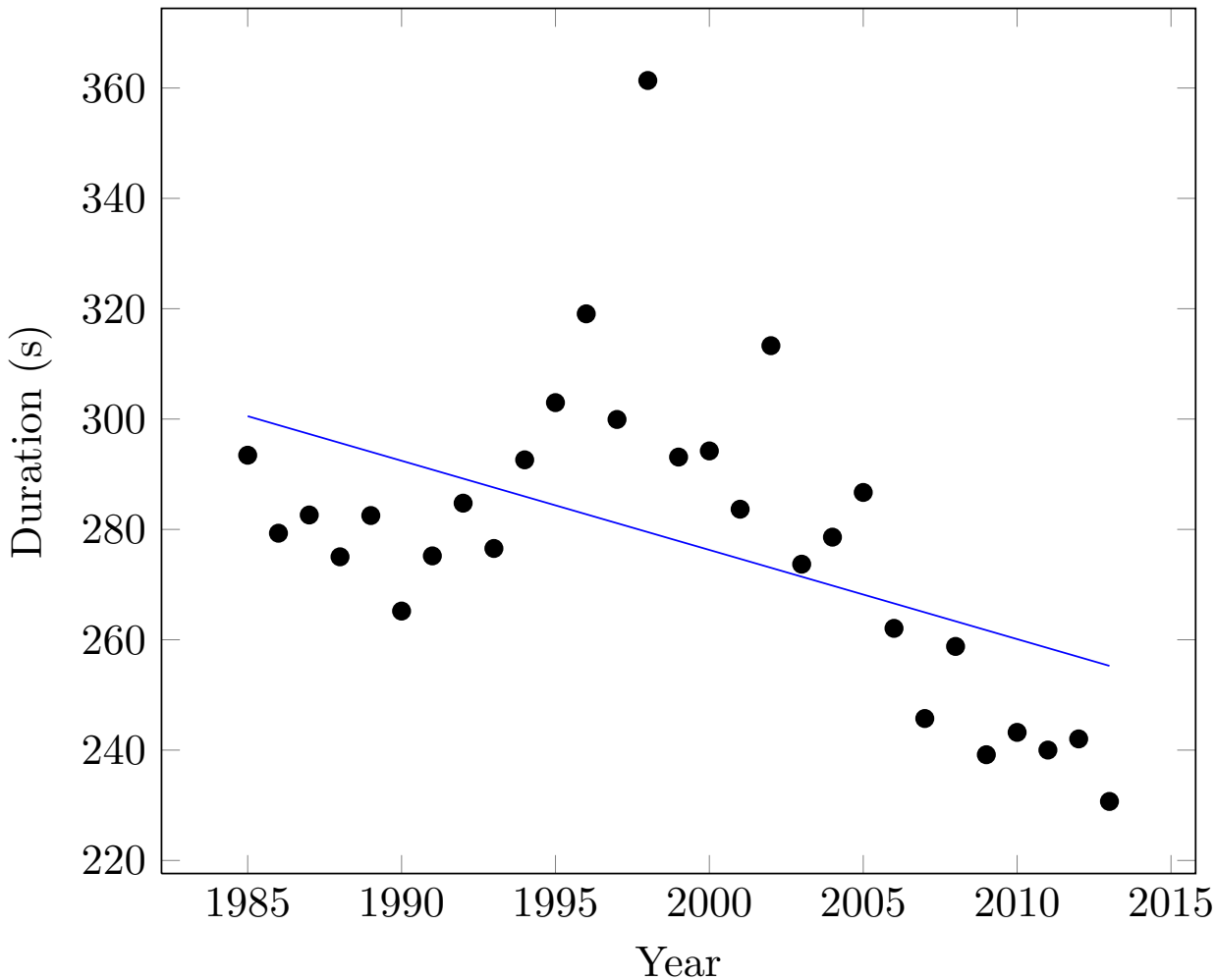
- Features with temporal aspect:
  - Time between beats
  - Timbre: 12 aspects (PCA)
    - Brightness (high vs low freq)
    - Flatness/narowness
    - Attack/sharpness
    - ...
- Statistical measures (calculated):
  - Median, mean, variance, stdev, min, max, range, 80th percentile
  - Skewness (3rd moment) (lean of the distribution)
  - Kurtosis (4th moment) (peakedness of the distribution)

Time Machine

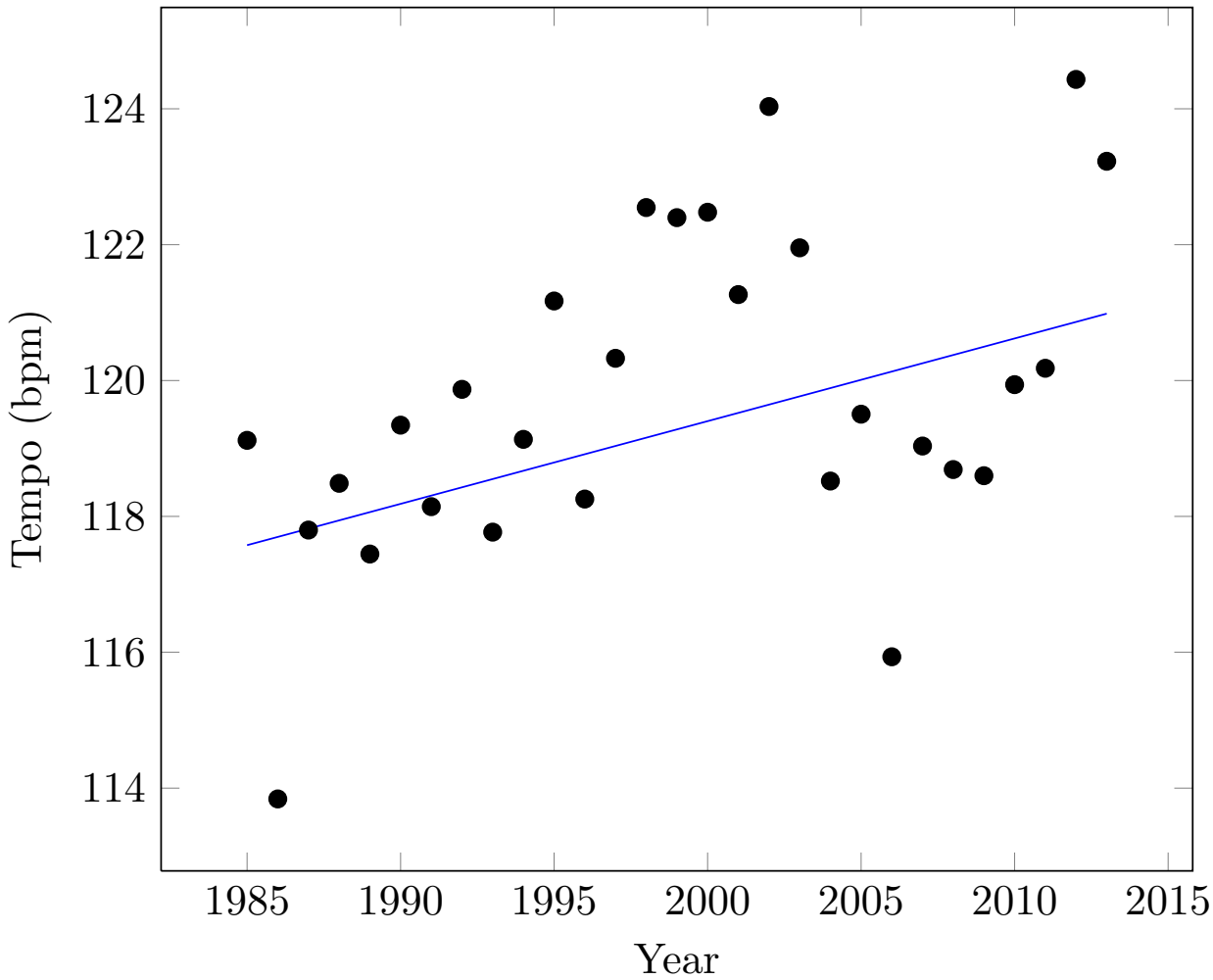


→ Billboard dataset

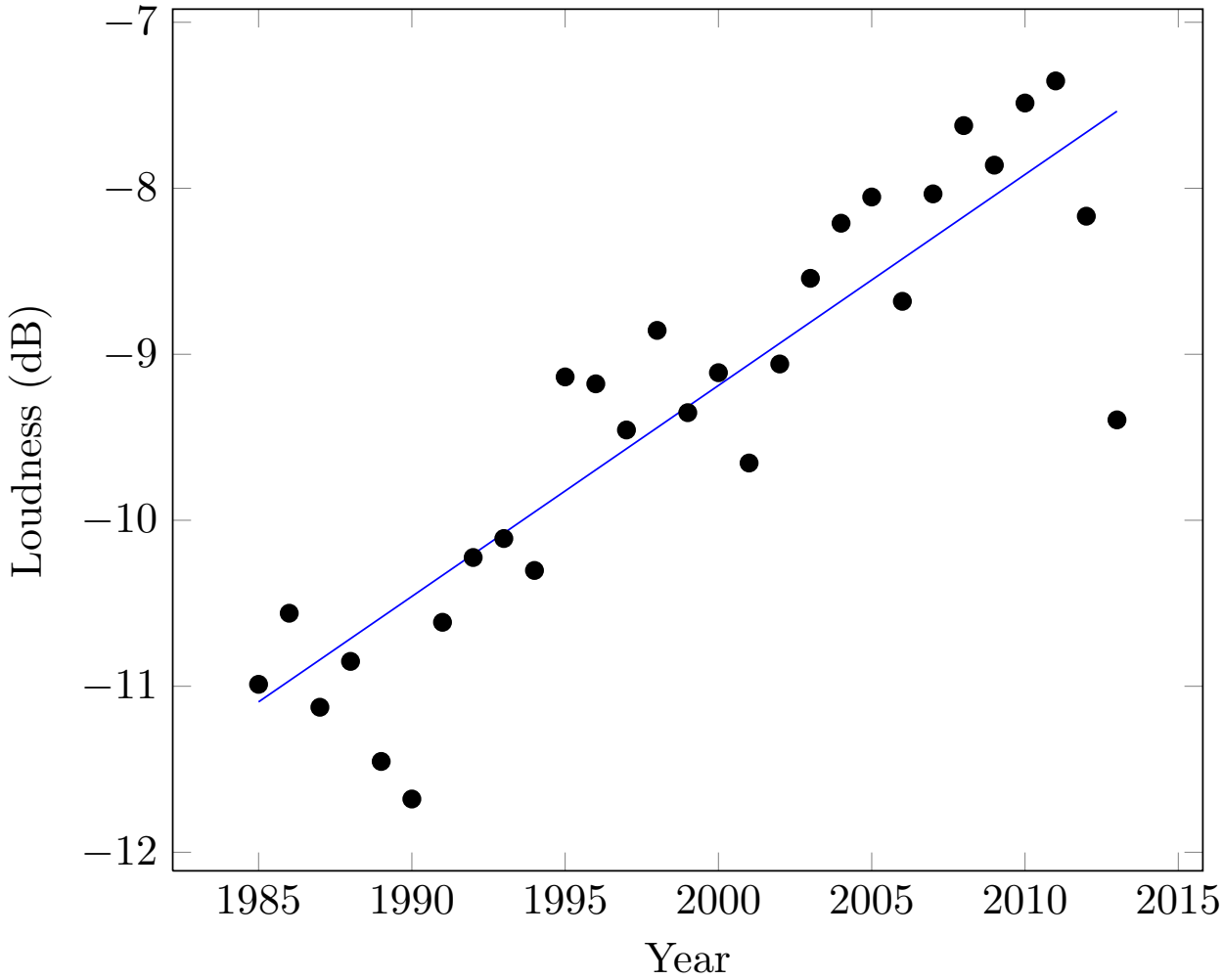
Duration


Duration (s)
360
340
320
300
280
260
240
220
1985
1990
1995
2000
2005
2010
2015
Year

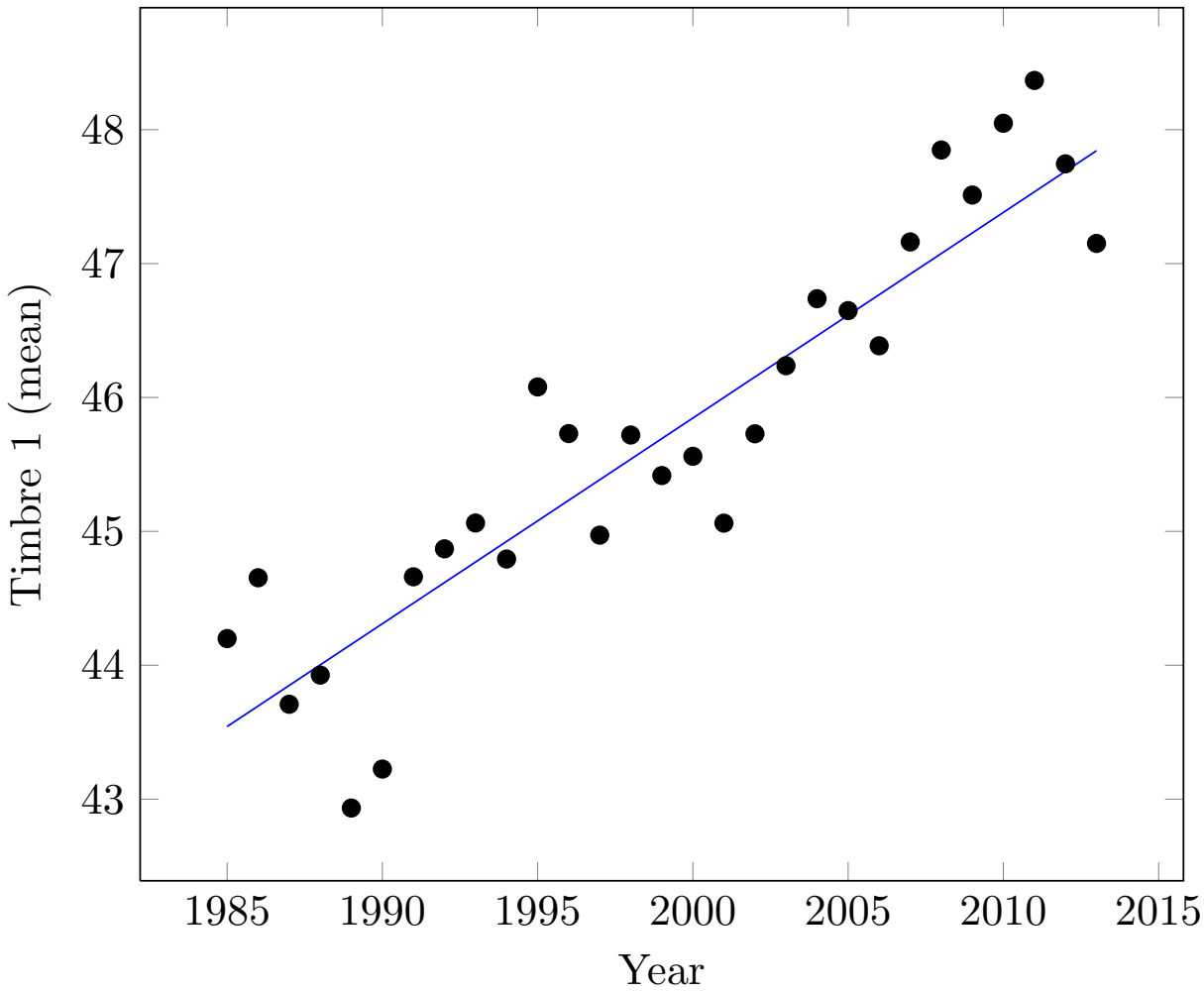
Tempo


Tempo (bpm)
124
122
120
118
116
114
1985
1990
1995
2000
2005
2010
2015
Year

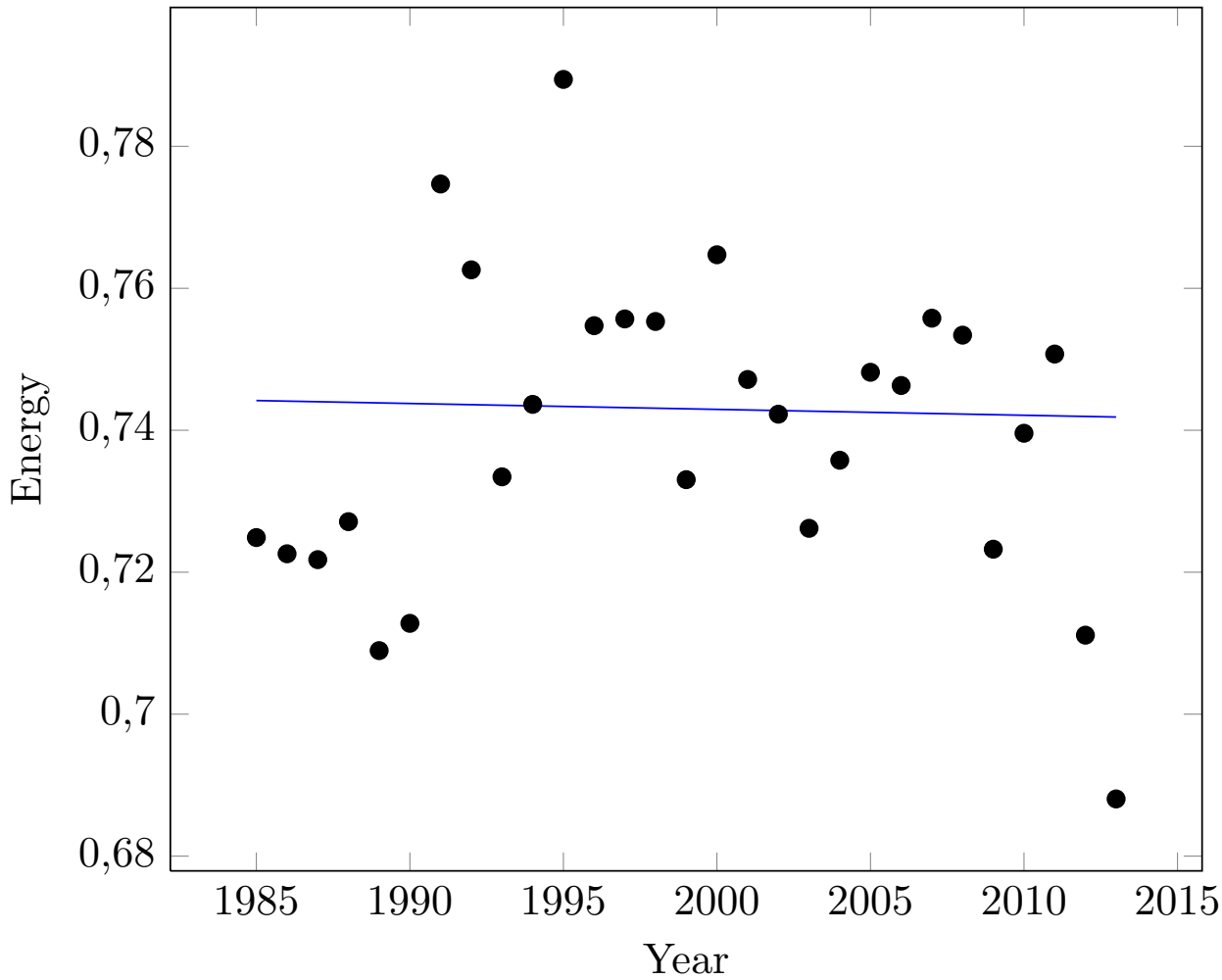
Loudness


−7
−8
−9
−10
−11
−12
Loudness (dB)
1985
1990
1995
2000
2005
2010
2015
Year

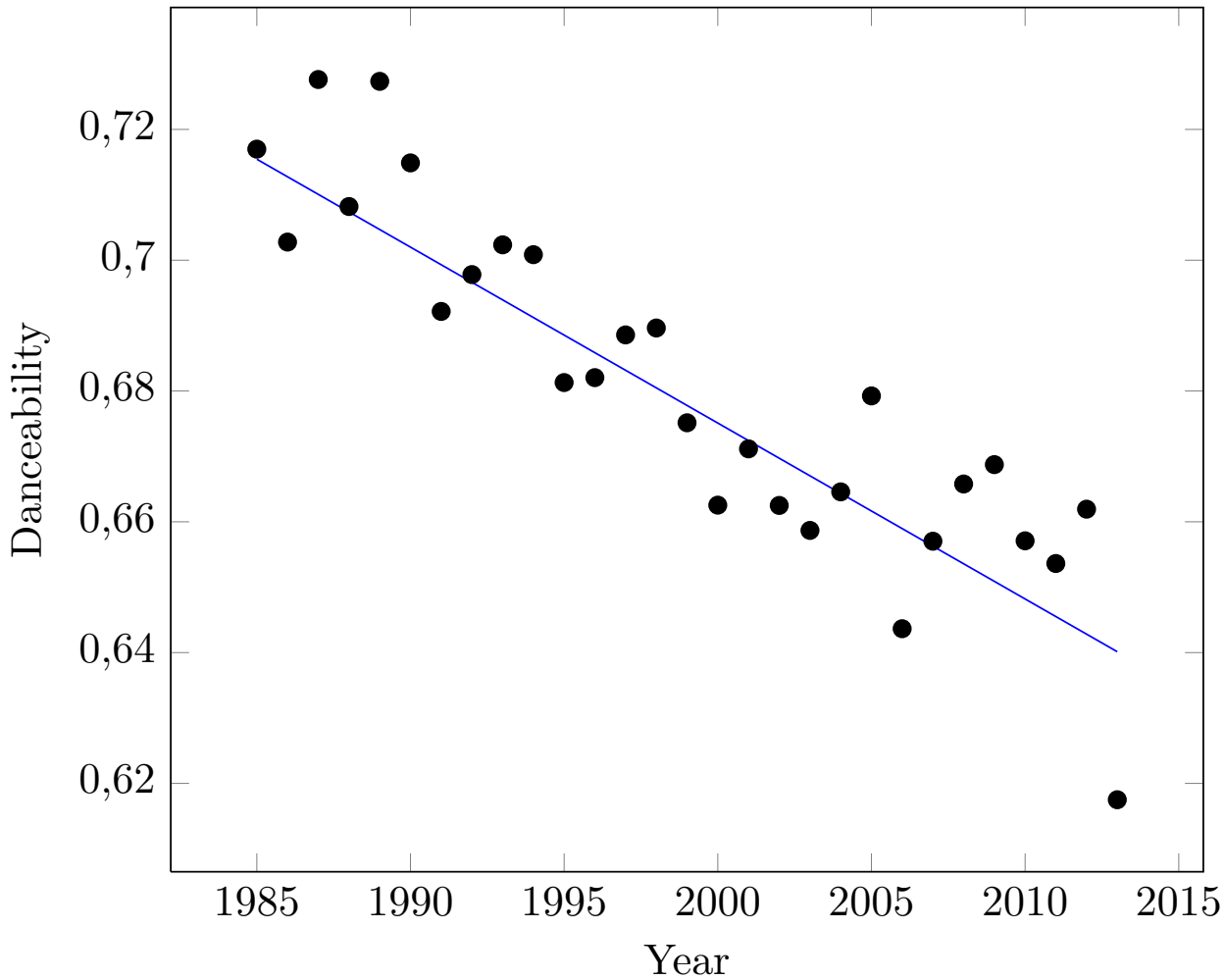
Timbre 1 (mean)



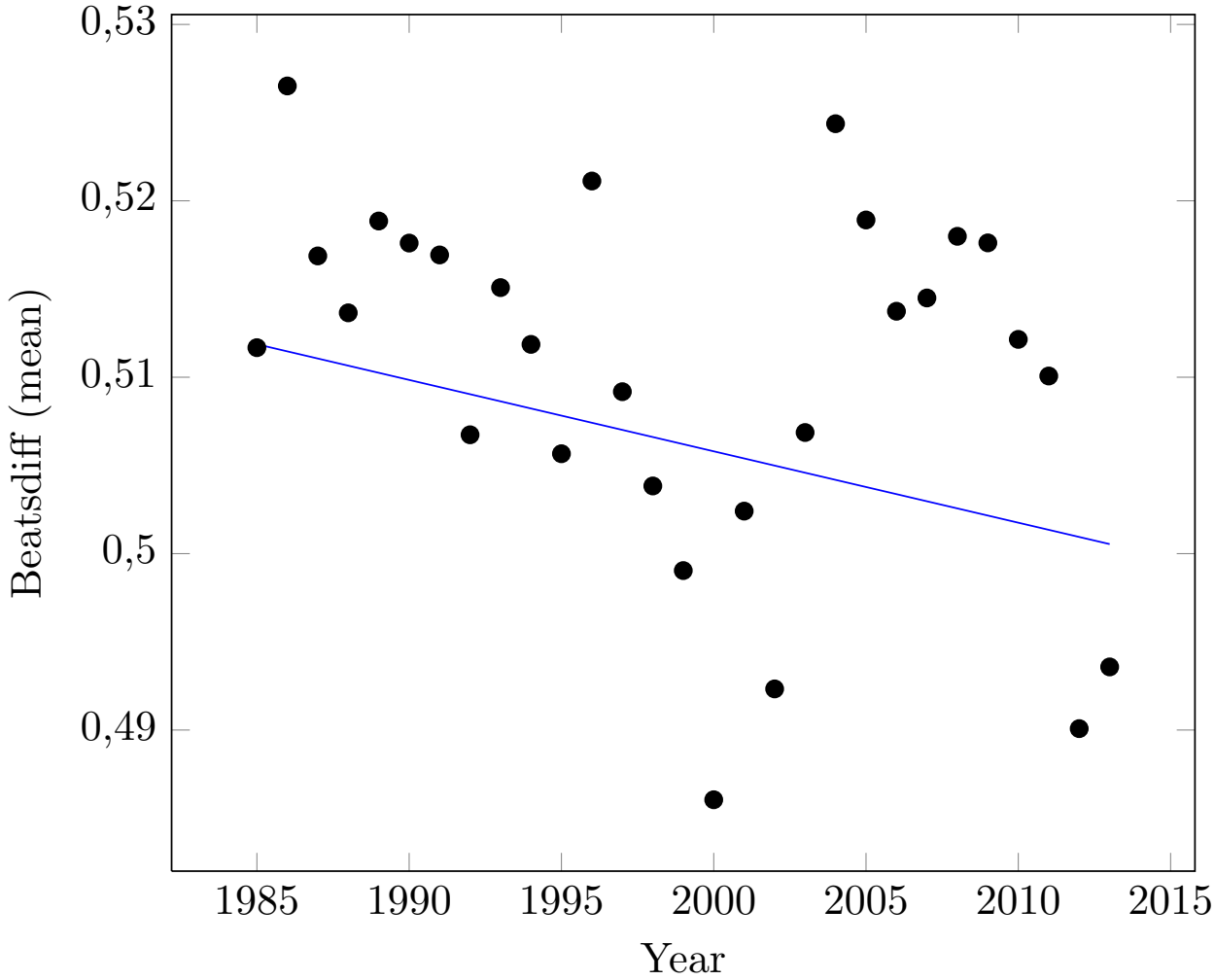
Energy


0,78
0,76
0,74
0,72
0,7
0,68
Energy
1985
1990
1995
2000
2005
2010
2015
Year

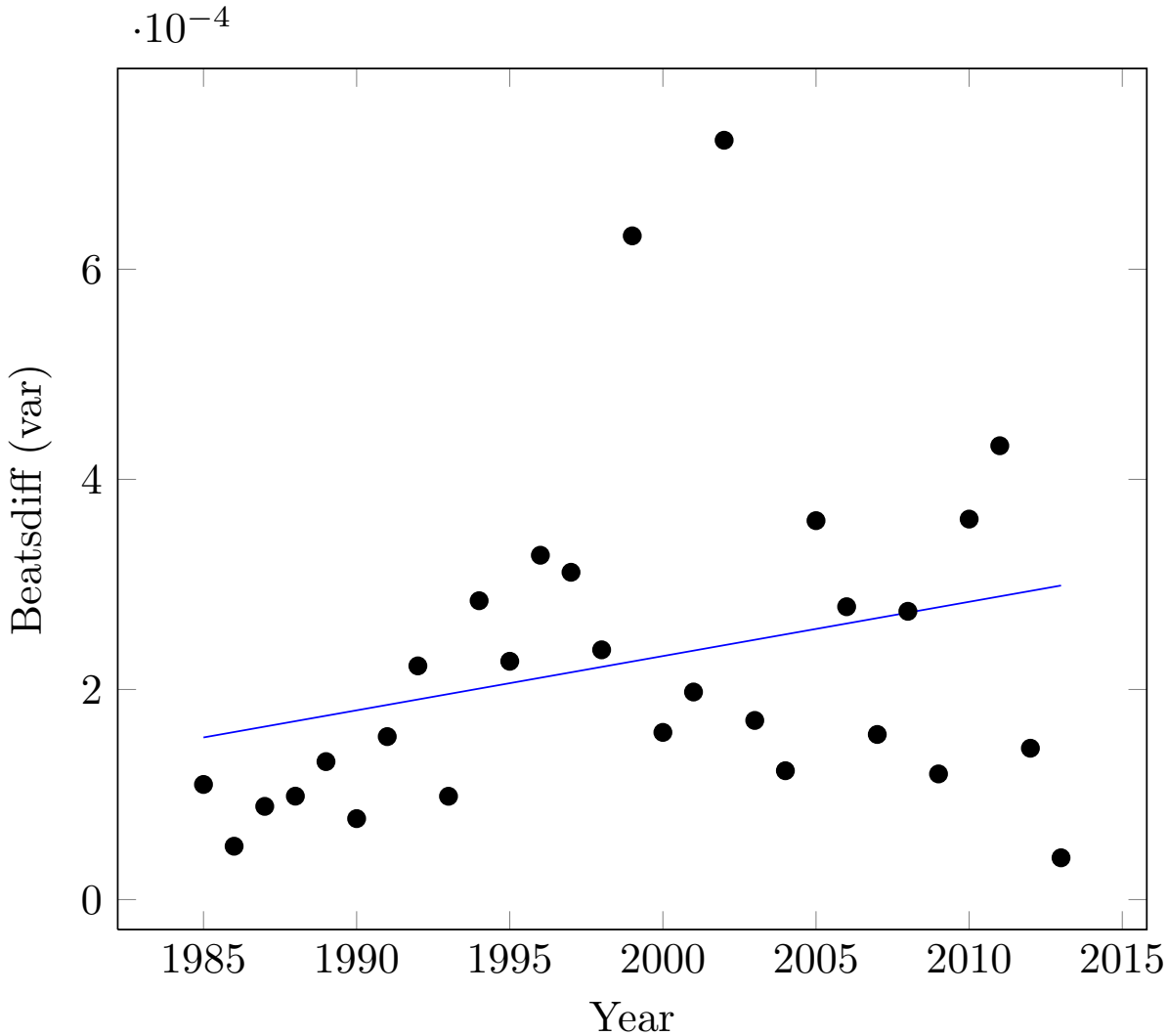
Danceability



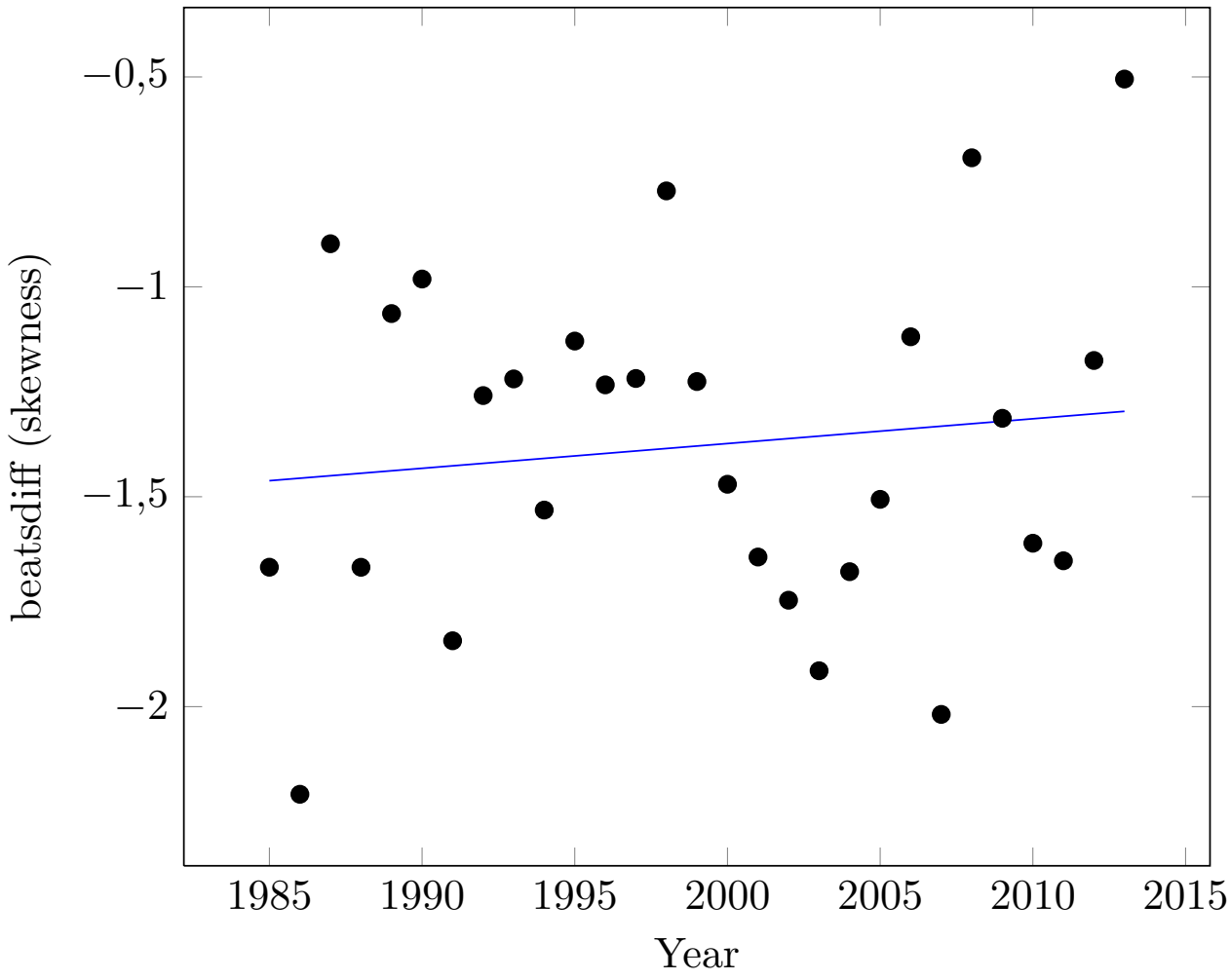
Beats Differences (mean)



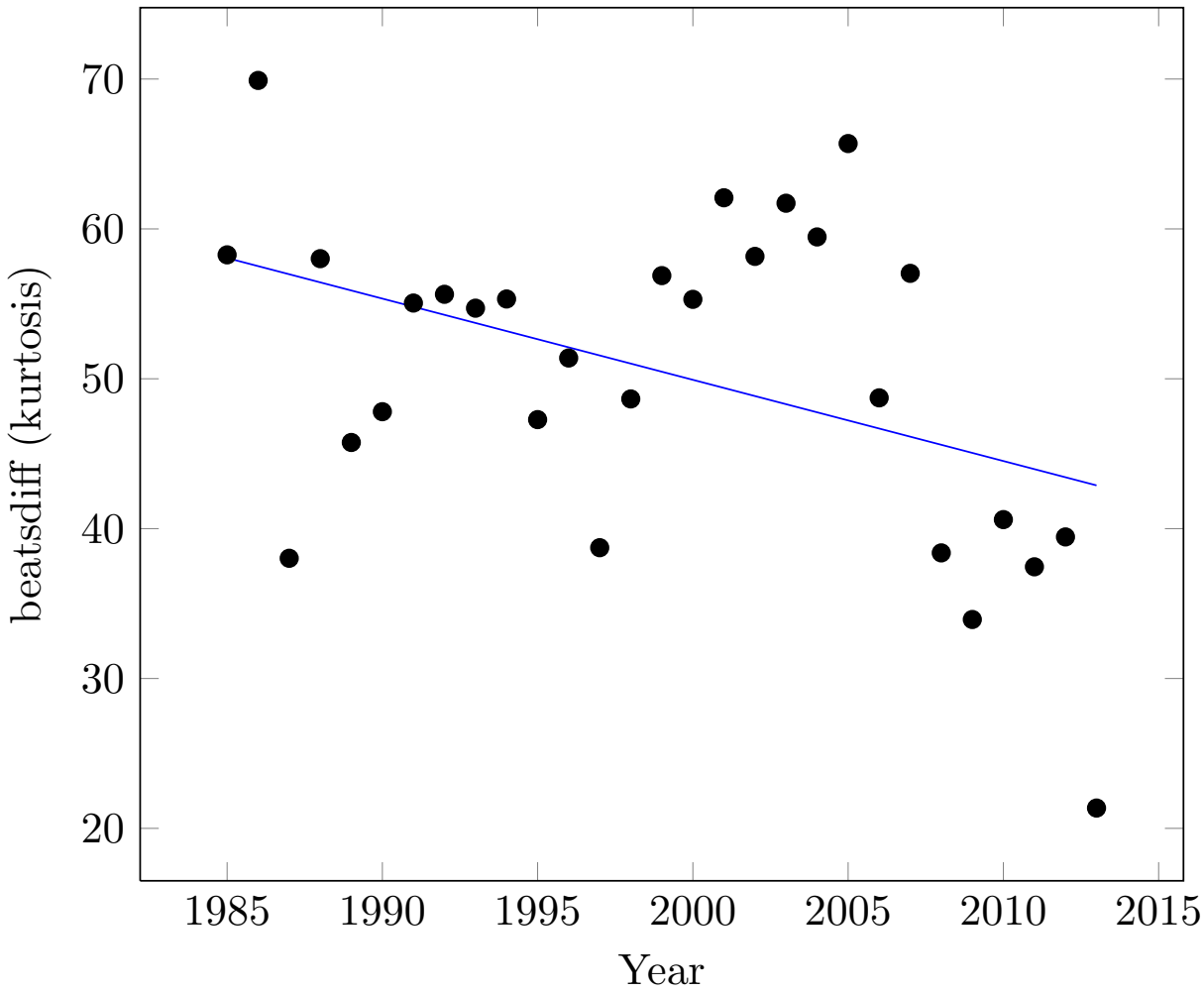
Beats Differences (var)



Beats Differences (skewness)



Beats Differences (kurtosis)



Hit prediction

- Hit versus non-hit:
  - Top 10 versus Top 30-40 (D1)
  - Top 10 versus Top 20-40 (D2)
  - Top 20 versus Top 20-40 (D3)
- 2009 until 2013
- Official Charts Company dataset
- 10-fold cross validation
- Weka Experimenter/Explorer

Input data

- Normalized (statistically): $x_n = \frac{x-\mu}{\sigma}$
- Input selection
  - Curse of dimensionality
  - CfsSubsetEval:
    - Individual predictive value
    - Degree of redundancy
  - With GeneticSearch
  - Result: 35–50 attributes

5 techniques

- Comprehensability:
  - C4.5 Tree (J48): divide-and-conquer approach
    $\rightarrow$ recusively, best separating feature, pruned
  - RIPPER Ruleset (JRip): sequential covering
    $\rightarrow$ one rule, covered instances removed, repeat

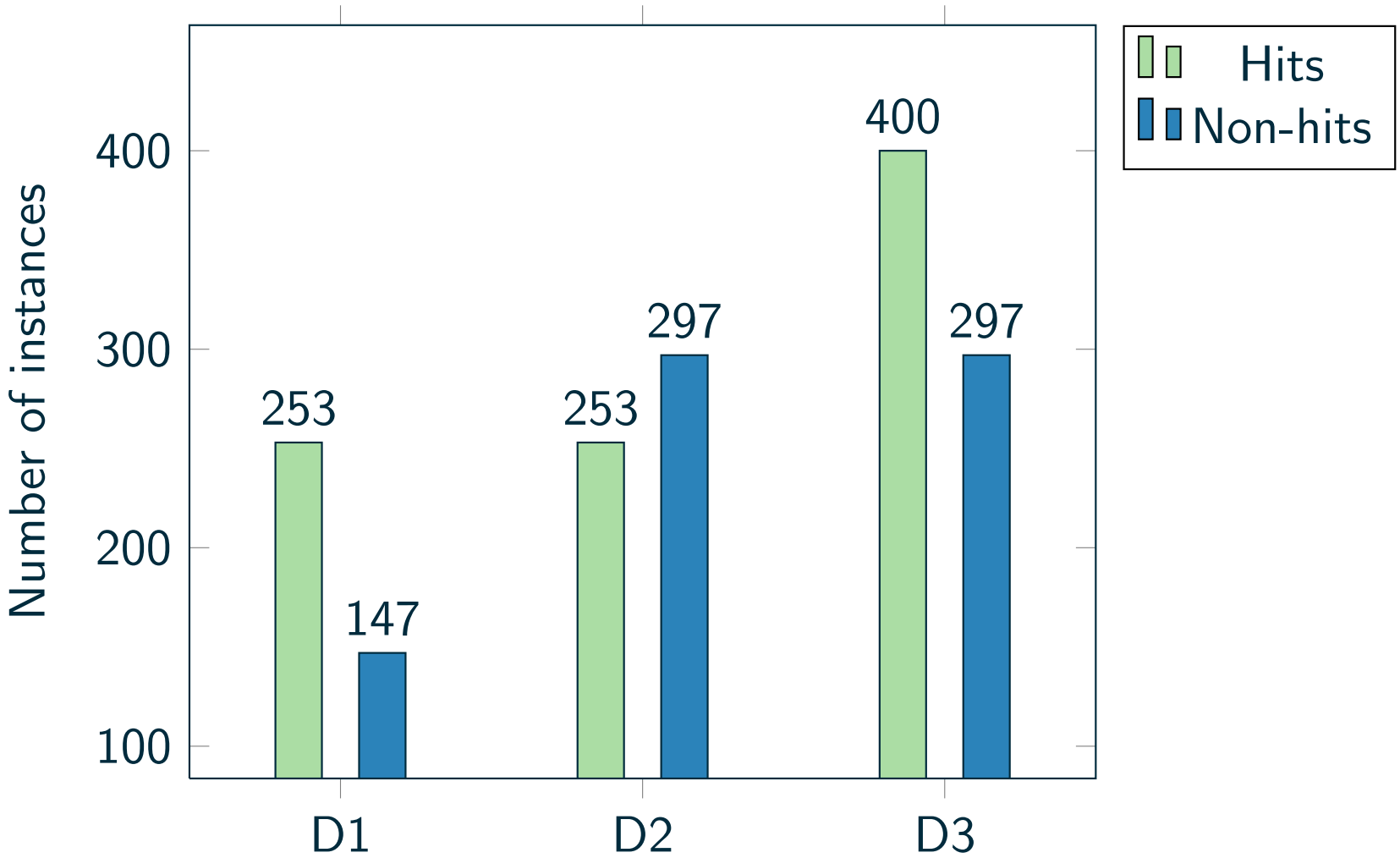
5 techniques

- Prediction:
  - SimpleLogistic: linear logistic regression model
  - NaiveBayes: estimates class-probablility based on assumption that attributes are conditionally independent.
  - SMO (SVM): sequential minimal optimization algorithm
    $\rightarrow$ Polynomial and RBF kernel

SVM parameter optimization

- CVParameterSelection
- GridSearch:
  - Multiple parameters:
    - c: 1–21 (+2)
    - gamma (RBF): 0.00001–10 (*10)
    - exponent (Poly): 1–2 (+1)
  - Weighted AUC (by class size)
  - Better results

Datasets


Number of instances
400
300
200
100
253
147
253
297
400
297
D1
D2
D3
Hits
Non-hits

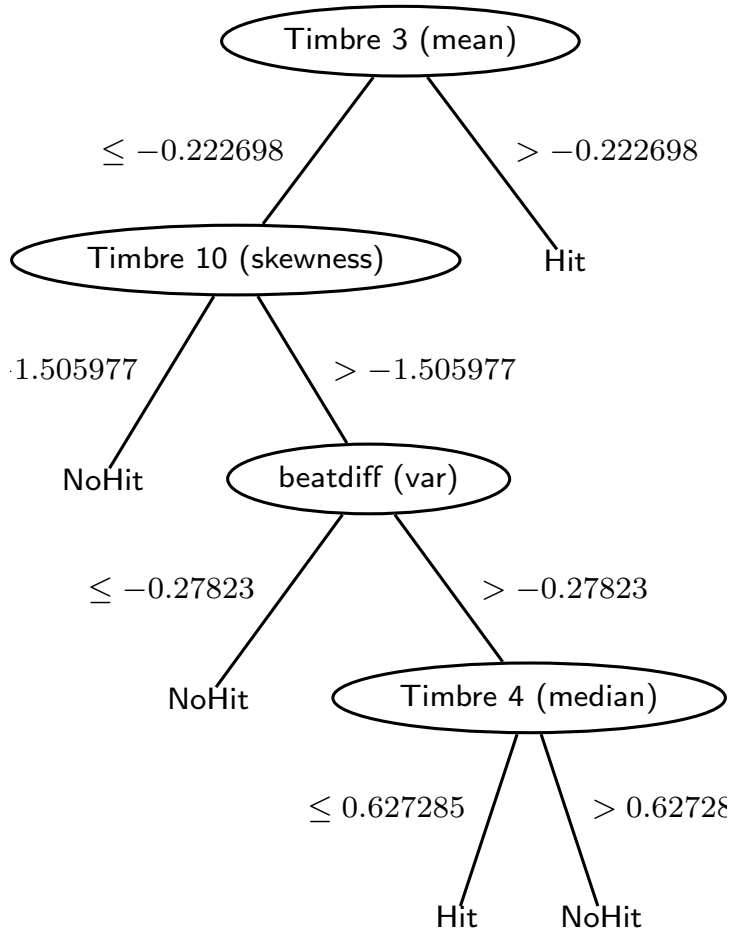
Results

Table : Results with 10-fold validation (AUC).

AUC	D1		D2		D3	
	-	IS	-	IS	-	IS
C4.5	*0.51*	0.58	*0.56*	0.55	0.52	*0.55*
RIPPER	*0.54*	*0.6*	0.58	*0.55*	0.54	0.54
Naive Bayes	**0.64**	**0.67**	**0.64**	**0.65**	**0.61**	0.62
Logistic regression	**0.65**	**0.68**	**0.66**	**0.63**	**0.64**	**0.65**
SVM (Polynomial)	0.57	**0.6**	**0.65**	0.6	0.57	**0.61**
SVM (RBF)	0.58	*0.6*	**0.63**	0.6	0.57	**0.6**

IS = input selection, $p < 0.01$: italic, $p > 0.05$: bold.

C4.5 Tree



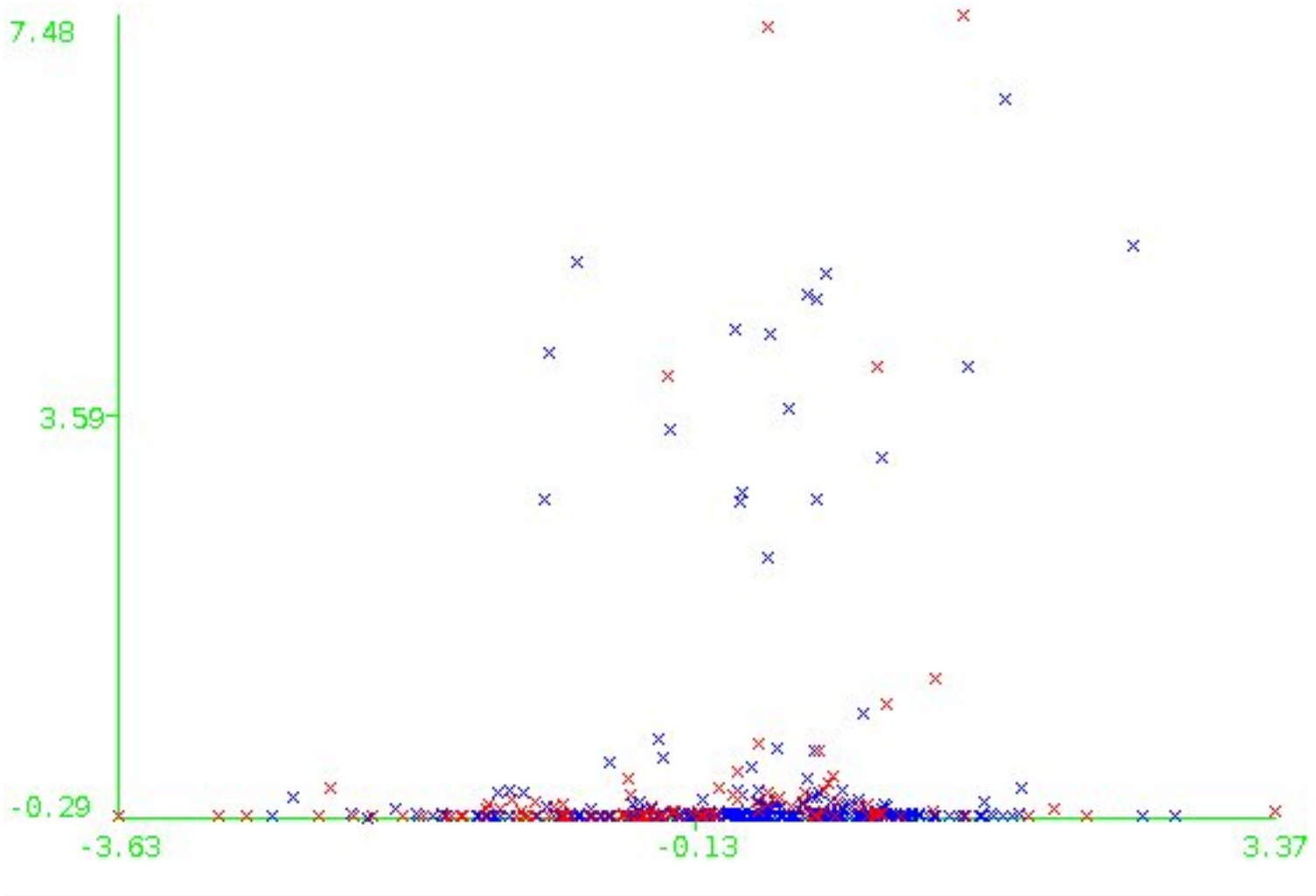
AUC: 0.58

RIPPER Ruleset (JRip)

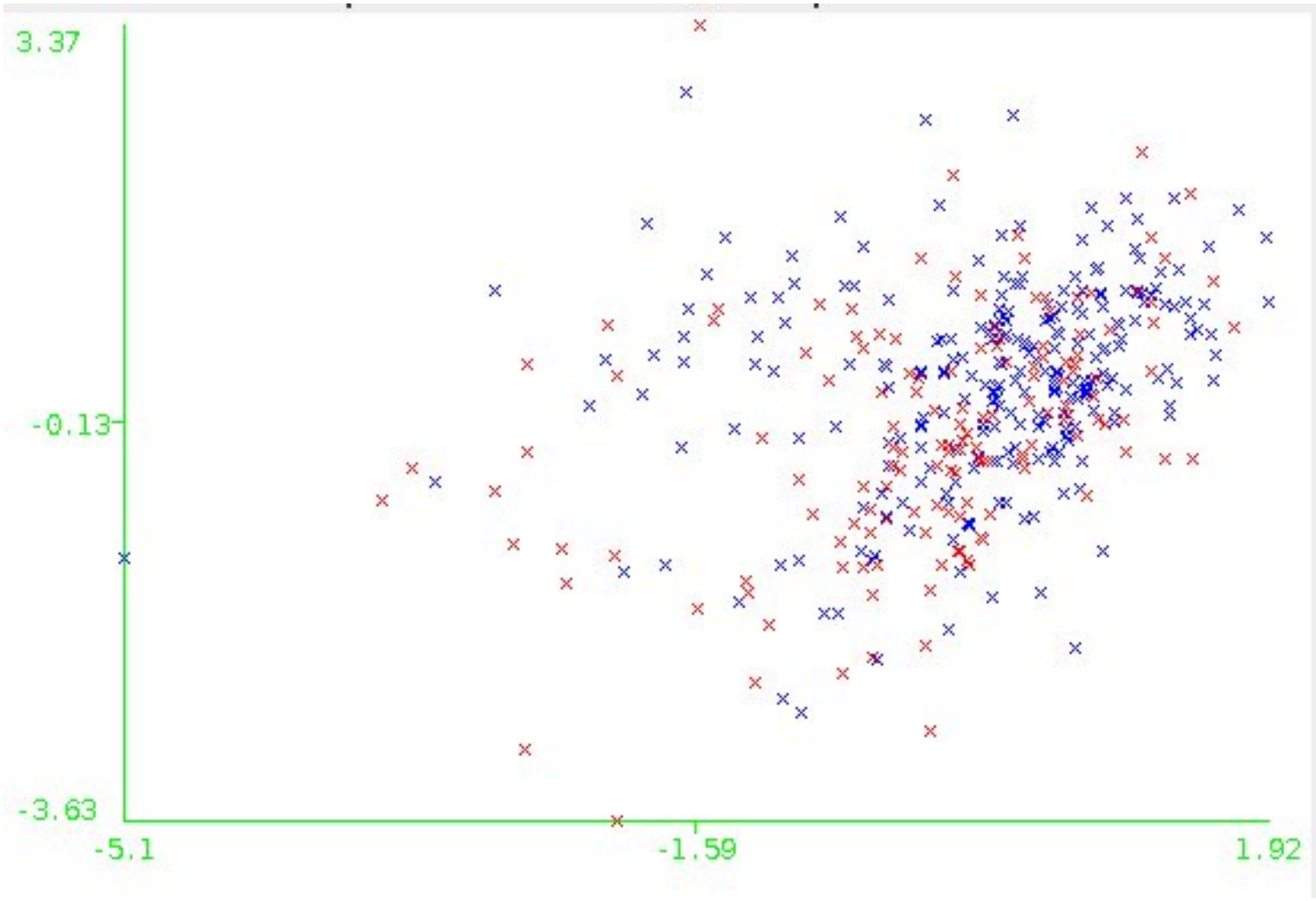
- (T3mean $\leq$ -0.356686) and (beatdifvariance $\leq$ -0.27823) $\Rightarrow$ NoHit
- (T3min $\leq$ -0.068467) and (T480perc $\geq$ 0.92738) $\Rightarrow$ NoHit
- $\Rightarrow$ Hit

AUC = 0.58

T3mean (x) versus beatdif var (y)



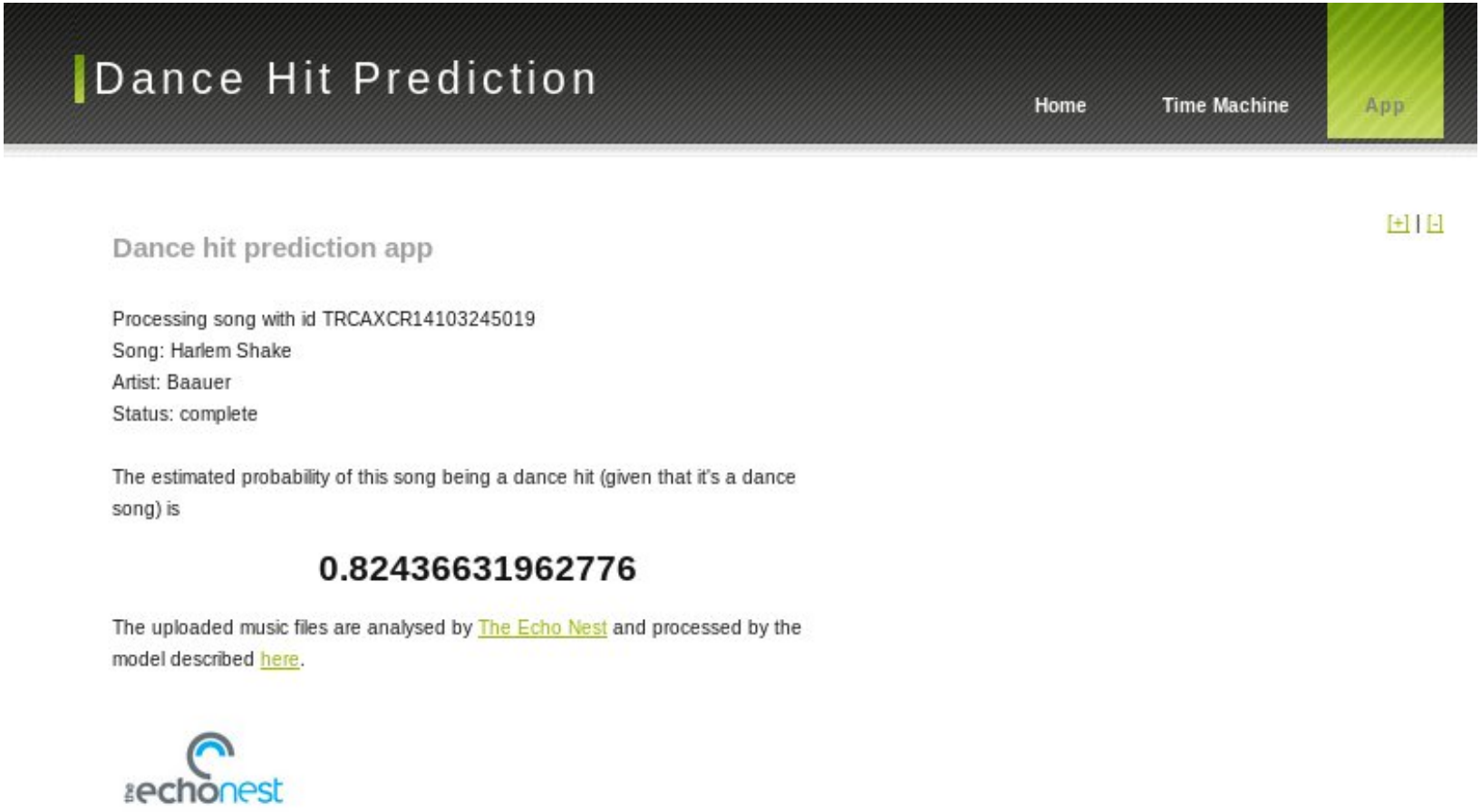
T1 median (x) vs T3 mean (y)



Out-of-time test set

	AUC		accuracy (%)	
	split	10CV	split	10CV
C4.5	0.603	0.58	62.5	63
RIPPER	0.466	0.6	62.5	65.5
Naive Bayes	0.777	0.67	75	**66.5**
Logistic regression	**0.794**	**0.68**	**80**	64.25
SVM (Polynomial)	0.700	0.6	62.5	64.25
SVM (RBF)	0.771	0.6	75	65

http://antor.ua.ac.be/dance



Conclusion

Multiple models were built that can predict if a dance song is going te be a top 10 hit and implemented in an online application.

Future research:

- More data (social network, lyrics, meta data, other music data)
- Different types of music
- Generate dance hits?

University of Antwerp
Operations Research Group
ANT/OR

Dance Hit Prediction

Dorien Herremans, David Martens, Kenneth Sörensen

Workshop on Music and Machine Learning
Prague, 23.09.2013

dorien.herremans@uantwerpen.be